Online Adspace Posts' Category Classification

Dhawal Joharapurkar
Manipal Institute of Technology
dmjan21@gmail.com

Vaishak Salin
Microsoft India Pvt. Ltd.
vaishak93@gmail.com

Vishal Krishna
Microsoft India Pvt. Ltd.
viscrisn@gmail.com

Abstract

Online adspaces require the seller/buyer to select a category to post their advertisements. This practice not only causes hindrance to legitimate users while posting their advertisements, but, also deter the user experience as they will see a lot of non-categorized ads due to human error or spam. Classifying advertisements posted on an online adspace can help in spam detection, better information propagation which in turn enhances user experience.

Craigslist is a prevalent platform for local classified advertisements. In this paper, we present a classification system for an online advertisement space such as Craigslist. We show the performance of our algorithm with three standard classifiers, *viz.*, Support Vector Machine, Random Forest, and Multinomial Naive Bayes. An accuracy of 84% was achieved with the SVM.

1 Introduction

With the advent of online adspaces, there are several hundred advertisements that get posted every second and such an influx of unstructured information is a rich source of data. Classification of advertisements into categories based on the text of the advertisement, a simple but rather useful technique is presented in this paper along with its implications.

Craigslist is an online place with local classifieds and forums which are community moderated, and largely free, on commodities like jobs, housing, goods, services, romance, local activities, advice, etc. There are different sub-domains of Craigslist, for each country they operate in. Each country has listings organized by sections which are then further subdivided into categories. We chose Craigslist as they have made some of their data publicly available is one of the largest online adspaces currently operating as shown by the statistics in the table below. However, this method can be applied to any online adpsace.

	users	Countries	posts/month
Craigslist	500M+	70+	80M+
Ebay	300M+	40	50M+

Table 1: Craigslist vs Ebay

As the table indicates, Craigslist is a more data rich source of advertisements. One reason for this is that Craigslist doesn't limit itself to goods only whereas other e-commerce websites do.

We use supervised learning techniques to classify a Craigslist post into different categories. We use text mining techniques like text normalization(Sproat et al., 2001) and term frequency inverse document frequency for preparing the data for classification. We tested several classification techniques and tabulated the results obtained using various assessment methods. SVM(Joachims, 1998) with a linear kernel yields the best result with an accuracy of 84% on the test data.

An online adspace such as Craigslist could use a recommender system that predicts the category under which the post must be filed while a user is posting an advertisement, using our classifier in an online setting. This sort of recommendation system is a very nifty way of improving the user experience for a site which sees on an average an excess of 80 million advertisements posted per month.

Our Contribution: We propose a classifier wherein, when a user is posting an advertisement on an online adspace such as Craigslist, a recom-

mendation is made to the user to select the category for the advertisement to be displayed in. In the case that the user doesn't select it, the post will be displayed in the predicted category by default.

2 Motivation

Recommendations can be made in an online setting wherein the text from the description of the advertisement is used to create feature vectors and fed into a classifier which predicts the category of the post. Based on this categorization, a recommendation is made to the user. This elementary step reduces the work of the user having to manually select the category to post his/her advertisement to. The category predicted can also be helpful in preventing the user from spamming the adspace by placing totally irrelevant ads to a particular category. This in turn enhances the user experience of the people browsing the advertisements listed.

Specifically, given the city, section and heading of a Craigslist post, we have to predict the category under which it should be posted.

3 Related Work

To our knowledge, there has been no work done in detecting spam (wrongly located advertisements) on online advertisement spaces. However, there has been some work in prediction of the price of an object using the title and the textual description of the advertisement by fitting a Naive Bayes classifier. (Elridge)

We chose to build such a classifier on Craigslist data as Fuxman et al. (Fuxman et al., 2009) have showed Craigslist to be a data source to improve classification accuracy in cases where a simple algorithm can outperform a sophisticated algorithm if it is provided with more training data.

4 Data

The dataset is a subset of the data generated on the Craigslist sites for a set of sixteen different cities (such as New York, Mumbai, etc.). The dataset has records from four sections forsale, housing, community and services and a total of sixteen categories from those sections. The categories are: activities, appliances, artists, automotive, cellphones, childcare, general, household-services, housing, photography, real-estate, shared, temporary, therapeutic, video-games and wanted-housing. Each category belongs to only one section. This data was made publicly available.

The first line in the dataset is an integer N. N lines follow, each line being a valid JSON object. The following fields of raw data, given in JSON:

- city (string, ASCII) : The city for which this Craigslist post was made
- section (string, ASCII) : for-sale / housing / community / services
- heading (string, UTF-8) : The heading of the post

A sample record looks as follows:

{***"city": "singapore", "section": "for-sale", "heading": "Panasonic ,2doors fridge(238L)($220 with delivery+1mth warranty)"***}

A total of approximately 20,000 records were made available, proportionally represented across these sections, categories and cities. The format of training data is the same as input format but with an additional field category, the category in which the post was made. A separate test dataset of 15,370 records was also provided, which we have used to test our final model on.

5 Experimental Setup

In our work, we have used scikit-learn(Pedregosa et al., 2011) (formerly scikits.learn) which is an open source machine learning library for the Python programming language. It features various classification, regression and clustering algorithms such as support vector machines, logistic regression, naive Bayes, random forests, gradient boosting, kmeans clustering and DBSCAN, etc. It is designed to interoperate with the Python numerical and scientific library SciPy (Jones et al., 2001).

5.1 Preprocessing of Data

Most category based ads have similar words. For example, words like BHK, wooden flooring, etc occur in house advertisements; MB, GB, camera, appear in cellphone advertisements.
One major feature of advertisements which are usually considered spam is the presence of lot of special characters. Most spam advertisements

	Accuracy	F1	Precision	Recall
SVM	0.81	0.82	0.83	0.82
Random Forest	0.73	0.75	0.76	0.75
Multinomial NB	0.68	0.74	0.75	0.74

Table 2: Baseline Performance of Algorithms

	Accuracy	F1	Precision	Recall
SVM	0.84	0.84	0.85	0.84
Random Forest	0.75	0.77	0.77	0.77
Multinomial NB	0.68	0.77	0.77	0.78

Table 3: Performance of Algorithms after Parameter Tweaking

have several special characters more than legitimate advertisements.

Feature generation: We remove all the special characters from the data as these arent really helpful from a lexical perspective, but use the count of these characters as a feature.

Normalization: We treat all numbers that occur equally as these usually represent the price or quantity of the product, which doesnt really help us decide the category of an advertisement. We normalize it by using the word numbr for any digit or number that occurs.

5.2 Creation of Classifier

The preprocessed data is then used to the create feature matrix. We use the bag of words model and then apply tf-idf (Salton and Yang, 1973) to generate feature vectors for each advertisement.

The feature matrix is sent to the classifier whose parameters have been optimized by the GridSearchCV module present in sklearn.(Pedregosa et al., 2011)

6 Evaluations and Results

Baseline results: The baseline results are calculated with the default/standard classifiers with no parameter optimizations done. The scores are averaged over 3-fold cross validation.

Results: After tweaking the parameters of the classifier using GridSearchCV our results improved and have been listed in Table 3.

From the confusion matrix, it is evident that the class labelled "therapeutic" was the best classified class with more than 1400 of the 1600 data points being correctly classified as "therapeutic".

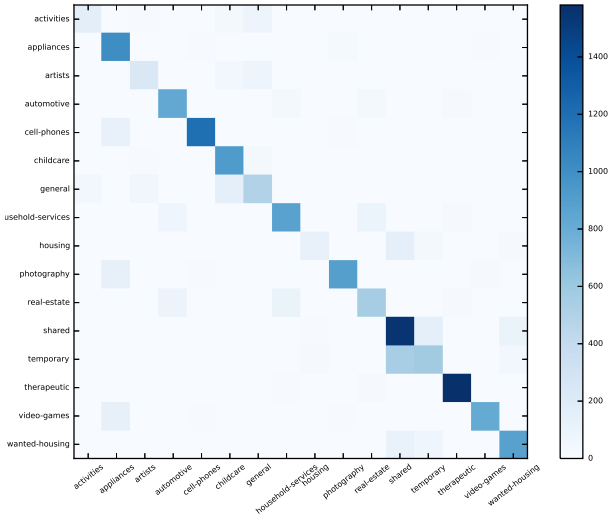

Figure 1: Confusion Matrix

This is an important result as it is well known that advertisements which fall under this category are most likely spam when found in other categories. Hence, classification on the text of the advertisement can help detect spam and block it out. We can also note for example, appliances get confused with cell phones and video games which is an interesting finding since the two are semantically closer than, say, housing.

Another stark importance of such a system is the amount of man hours saved. With 80M posts coming in every month, let us say a user spends about 30 seconds selecting the section and category of the advertisement. That is approximately a whopping 666,000+ man hours saved each month.

7 Conclusion

In this paper, we present a classifier for online adspaces, which classifies an advertisement amongst 16 categories. If an advertisement is accurately classified in its correct class, it helps increase the reach of the advertisement as it is displayed to the people looking in the particular section and is not classified as spam. In this work,

we have used the Craigslist data which is publicly available, and, very rich in terms of categories and datapoints. Our classifier secures an mean accuracy of 84% when tested using cross-validation.

8 Future Work

We look to improve this work in two ways. First, we would like to expand the current classifier to be a multi-class classifier. This would help classify ads which can belong to more than one category, thereby it would result in the advertisement being displayed in more than one category and hence reach more "prospective" buyers. As an alternative, it would be interesting to treat this as a multi-label classification problem, wherein each item can be labelled with more than one category. This allows us to tag each advertisement with classes which are related.

Secondly, we would want to explore and implement better feature engineering. The section-category tags of posts follow a hierarchy. Being able to incorporate this hierarchy as features would, we believe, significantly improve our feature vectors. Using such a classifier in setting where advertisements can comprise a code-mixed dataset like on Indian online adspaces such as Olx, Junglee, etc., can will benefit from learning features from such data.

References

Ariel Fuxman, Anitha Kannan, Andrew B Goldberg, Rakesh Agrawal, Panayiotis Tsaparas, and John Shafer. 2009. Improving classification accuracy using automatically extracted training data. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1145–1154. ACM.

Thorsten Joachims. 1998. *Text categorization with support vector machines: Learning with many relevant features*. Springer.

Eric Jones, Travis Oliphant, Pearu Peterson, et al. 2001–. SciPy: Open source scientific tools for Python. [Online; accessed 2015-08-22].

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Gerard Salton and Chung-Shu Yang. 1973. On the specification of term values in automatic indexing. *Journal of documentation*, 29(4):351–372.

Richard Sproat, Alan W Black, Stanley Chen, Shankar Kumar, Mari Ostendorf, and Christopher Richards. 2001. Normalization of non-standard words. *Computer Speech & Language*, 15(3):287–333.